Origin and Phylogeny of Chloroplasts Revealed by a Simple Correlation Analysis of Complete Genomes

Ka Hou Chu[1], Ji Qi[2], Zu-Guo Yu[3,4] & Vo Anh[3]

[1]Department of Biology, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China

[2]Institute of Theoretical Physics, The Chinese Academy of Sciences, Beijing 100080, China

[3]Centre in Statistical Science and Industrial Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia.

[4]Department of Mathematics, Xiangtan University, Hunan 411105, China

Correspondence should be addressed to K.H. Chu.

Address: Department of Biology, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China. Tel: 852-26096772; Fax: 852-26035391; e-mail: kahouchu@cuhk.edu.hk.

Keywords: Chloroplast; genome; plant; phylogeny

Running head: Correlation analysis of chloroplast genomes

Abstract

The complete sequenced genomes of chloroplast have provided much information on the origin and evolution of this organelle. In this paper we attempted to use these sequences to test a novel approach for phylogenetic analysis of complete genomes based on correlation analysis of compositional vectors. All protein sequences from 21 complete chloroplast genomes were analyzed in comparison with selected archaea, eubacteria and eukaryotes. The distance-based analysis shows that the chloroplast genomes are most closely related to cyanobacteria, consistent with the endosymbiotic origin of chloroplasts. The chloroplast genomes are separated to two major clades corresponding to chlorophytes (green plants) *s.l.* and rhodophytes (red algae) *s.l.* The interrelationships among the chloroplasts are largely in agreement with the current understanding on chloroplast evolution. For instance, the analysis places the chloroplasts of two chromophytes (*Guillardia* and *Odontella*) within the rhodophyte lineage, supporting secondary endosymbiosis as the source of these chloroplasts. The relationships among the green algae and green plants in our tree also agree with results from traditional phylogenetic analyses. Thus this study establishes the value of our simple correlation analysis in elucidating the evolutionary relationships among genomes. It is hoped that this approach will provide insights on comparative genome analysis.

Introduction

Chloroplast DNA is a primary source of molecular variations for phylogenetic analysis of photosynthetic eukaryotes. During the past decade the availability of complete chloroplast genome sequences has provided a wealth of information to study the origin, including primary and secondary endosymbioses (Delwiche 1999; McFadden 2001a), and phylogeny of photosynthetic eukaryotes at the deep levels of evolution. There have been many phylogenetic analyses based on comparison of sequences of multiple protein-coding genes in chloroplast genomes (e.g., Martin et al. 1998, 2002; Turmel, Otis and Lemieux 1999, 2002; Adachi et al. 2000; Lemieux, Otis and Turmel 2000; De Las Rivas, Lozano and Ortiz 2002). Alternative methodologies for phylogenetic analysis of complete genomes have been proposed, for example, based on the rearrangement of gene order (Sankoff et al. 1992), the presence and absence of protein-coding gene families (Fitz-Gibbon and House 1999), gene content and overall similarity (Tekaia, Lazcano and Dujon 1999), and occurrence of folds and orthologs (Lin and Gerstein 2000). Yet the above approaches are all based on alignment of homologous sequences, and it is apparent that much information (such as gene rearrangement and insertions/deletions) in these data sets are lost after sequence alignment, let alone the intrinsic problems of alignment algorithms (Li et al. 2001; Stuart, Moffet and Baker 2002). There have been a number of recent attempts to develop methodologies that do not require sequence alignment for deriving species phylogeny based on overall similarities of the complete genome data (e.g., Li et al. 2001; Yu and Jiang 2001; Edwards et al. 2002; Stuart, Moffet and Baker 2002; Stuart, Moffet and Leader 2002). One of us (J. Qi) and his colleagues have developed a simple correlation analysis of

complete genome sequences based on compositional vectors without the need of sequence alignment. The compositional vectors calculated based on frequency of amino acid strings are converted to distance values for all taxa and the phylogenetic relationships were inferred from the distance matrix using conventional tree-building methods (see Materials and Methods for details). An analysis based on this method using 82 prokaryotes and 2 eukaryotes yielded a tree separating the three domains of life, Archaea, Eubacteria and Eukarya with the relationships among the taxa correlating with those based on traditional analyses (Qi, Wang and Hao, in press). A correlation analysis based on a different transformation of compositional vectors was recently reported by Stuart, Moffet and Baker (2002) and Stuart, Moffet and Leader (2002) who demonstrated the applicability of the method in revealing phylogeny using vertebrate mitochondrial genomes. In the present study we applied the above approach to analyze 21 complete chloroplast genomes, together with the genomes of 2 archaea, 8 eubacteria (including 2 cyanobacteria) and 3 eukaryotes (see Materials and Method for a list of complete nuclear and chloroplast genomes analyzed). The aim is to test the applicability of this correlation analysis in elucidating the origin and phylogeny of chloroplasts.

Materials and Methods

Genome Data Sets

Complete sequences of 21 chloroplast genomes (*Cyanophora paradoxa, Cyanidium caldarium, Porphyra purpurea, Guillardia theta, Odontella sinensis, Euglena gracilis, Chlorella vulgaris, Nephroselmis olivacea, Mesostigma viride, Chaetosphaeridium globosum, Marchantia polymorpha, Psilotum nudum, Pinus thunbergii, Oenothera elata,*

Lotus japonicus, Spinacia oleracea, Nicotiana tabacum, Arabidopsis thaliana, Oryza sativa, Triticum aestivu and *Zea mays*) and genomes of 2 archaea (*Archaeoglobus fulgidu* and *Sulfolobus solfataricus*), 8 eubacteria (*Helicobacter pylori, Neisseria meningitides, Rickettsia prowazekii, Borrelia burgdorferi, Chlamydophila pneumoniae, Mycobacterium leprae, Nostoc* sp. and *Synechocystis* sp.) and 3 eukaryotes (*Saccharomyces cerevisiae, Arabidopsis thaliana* and *Caenorhabitidis elegans*) were retrieved from NCBI database (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_euk_o.html).

Composition Vectors and Distance Matrix

We base our analysis on all protein sequences including hypothetical reading frames from each genome, regarding sequences of the 20 amino acids as symbolic sequences. In such a sequence of length L, there are a total of $N = 20^K$ possible types of strings of length K. We used a window of length K and slide it through the sequences by shifting one position at a time to determine the frequencies of each of the N kinds of strings in each genome. A protein sequence was excluded if its length is shorter than K. The observed frequency $p(\alpha_1\alpha_2...\alpha_K)$ of a K-string $\alpha_1\alpha_2...\alpha_K$ is defined as $p(\alpha_1\alpha_2...\alpha_K) = n(\alpha_1\alpha_2...\alpha_K)/(L-K+1)$, where $n(\alpha_1\alpha_2...\alpha_K)$ is the number of times that $\alpha_1\alpha_2...\alpha_K$ appears in this sequence. For example, in the protein sequence "MKRTFQPSILKRNRSHGFRIRMATKNGRYILSRRRAKLRTRLTVSSK", $p(R) = 11/47$, $p(MR) = 0$, $p(RR) = 2/(47-2+1) = 1/23$ and $p(RRR) = 1/(47-3+1) = 1/45$. Denoting by m the number of protein sequences from each complete genome, the observed frequency of a K-string $\alpha_1\alpha_2...\alpha_K$ is defined as $(\sum_{j=1}^{m} n_j(\alpha_1\alpha_2...\alpha_K))/(\sum_{j=1}^{m}(L_j-K+1))$; here

$n_j(\alpha_1\alpha_2...\alpha_K)$ means the number of times that $\alpha_1\alpha_2...\alpha_K$ appears in the *j*th protein sequence and L_j the length of the *j*th protein sequence in this complete genome.

Mutations occur in a random fashion at the molecular level, while selections shape the direction of evolution. There is always some randomness in the composition of protein sequences, revealed by statistical properties of protein sequences at single amino acid or oligopeptide level (see Weiss et al. 2000 for a recent discussion on this point). In order to highlight the selective diversification of sequence composition, we subtract the random background from the simple counting results. Supposed that we have performed direct counting for all strings of length $(K-1)$ and $(K-2)$, the expected frequency of appearance of K-strings is predicted by using a Markov model (Brendel, Beckman and Trifonov 1986):

$$q(\alpha_1\alpha_2...\alpha_K) = \frac{p(\alpha_1\alpha_2...\alpha_{K-1})p(\alpha_2\alpha_3...\alpha_K)}{p(\alpha_2\alpha_3...\alpha_{K-1})},$$

where q denotes the predicted frequency (when $p(\alpha_2\alpha_3...\alpha_{K-1}) = 0$, then definitely $p(\alpha_1\alpha_2...\alpha_{K-1}) = 0$ because a string will not appear if its sub-string does not appear; in this case we set $q(\alpha_1\alpha_2...\alpha_K) = 0$). In the above example, $q(RRR) = (1/23 \times 1/23)/(11/47)$. The above predictor via a Markov model has been used in biological sequence analyses (see Brendel et al. 1986 for example; see also Percus 2002, p. 47, for a theoretical development). A key step of our approach is to remove the above random background before performing a cross-correlation analysis (similar to removing a time-varying mean in time series before computing the cross-correlation of two time series). We then calculate a new measure *X* of the shaping role of selective evolution as

$$X(\alpha_1\alpha_2...\alpha_K)=\begin{cases} p(\alpha_1\alpha_2...\alpha_K)/q(\alpha_1\alpha_2...\alpha_K)-1, & if \quad q(\alpha_1\alpha_2...\alpha_K)\neq 0 \\ 0, & if \quad q(\alpha_1\alpha_2...\alpha_K)=0. \end{cases}$$

As an example, we display a segment of *p* for Chloroplast *Cyanophora paradoxa* in the left figure of Fig. 1 and the corresponding sequence *X* for the same set of *K*-strings in the right figure of Fig. 1. The transformation $X=(p/q)-1$ has the desired effect of removing the random background in *p* and rendering it a stationary time series suitable for subsequent cross-correlation analysis.

For all possible strings $\alpha_1\alpha_2...\alpha_K$, we used $X(\alpha_1\alpha_2...\alpha_K)$ as components to form a composition vector for a genome. To further simplify the notation, we used X_i for the i-th component corresponding to the string type i, $i = 1,\ldots, N$ (the *N* strings were arranged in a fixed order as the alphabetical order). Hence we constructed a composition vector $X=(X_1, X_2,..., X_N)$ for genome X, and likewise $Y=(Y_1, Y_2,..., Y_N)$ for genome Y.

If we view the *N* components in vectors *X* and *Y* as the samples of two zero-mean random variables respectively, the correlation $C(X,Y)$ between any two genomes X and Y is defined in the usual way in probability theory as $C(X,Y)=\dfrac{\sum_{i=1}^{N} X_i \times Y_i}{(\sum_{i=1}^{N} X_i^2 \times \sum_{i=1}^{N} Y_i^2)^{\frac{1}{2}}}$.

The distance $D(X,Y)$ between the two genomes is then defined as the equation $D(X,Y)=(1-C(X,Y))/2$.

Tree Construction and Statistical Test of the Trees

A distance matrix for all the genomes under study was generated. Different distance methods, including Fitch-Margoliash (Fitch and Margoliash 1967), neighbour-joining

(Saitou and Nei 1987) and minimum evolution (Saitou and Imanishi 1989), were then used to construct the phylogenetic trees.

Remark 1. The peptide frequency vector described in Stuart, Moffet and Baker (2002) and Stuart, Moffet and Leader (2002) is exactly the vector p that we described. However, their method of structure removal is entirely different from our method. Starting from the vector p, Stuart *et al*., (2002) used Singular Value Decomposition (SVD), then Dimension Reduction on their constructed matrix. The correlation distance is then used to construct the tree. In our method, we remove random background via a Markov model for q and X. The SVD step is much more complicated than our method in both theoretical and practical sense.

A previous study on prokaryotes shows that the topology of the trees stabilized for $K \geq 5$ (Qi, Wang and Hao, in press). In the present study, we used $K = 4$ or 5 in our analysis and the topologies of the resulting trees are similar. Here we present the results based on $K= 5$. We conducted the analysis on all the 34 genomes, as well as on the 21 chloroplast genomes alone using *Synechocystis* as the outgroup. The former analysis aims to explore the origin of the chloroplast genome whereas the latter analysis is for comparison with previous phylogenetic analyses (Martin et al. 1998, 2002; Turmel, Otis and Lemieux 1999; De Las Rivas, Lozano and Oritz 2002) that include most of chloroplast genomes as in our analysis using the same outgroup taxon. The distance matrix generated from this analysis is available at http://www.itp.ac.cn/~qiji.

Bootstrapping was performed to give statistical support to the phylogenetic trees. Sequences of proteins were drawn randomly from a complete genome until the total number of proteins selected in each bootstrap was equal to the number of protein-coding

genes of that particular genome. That is, in each bootstrap some proteins might be selected more than once while others might not be included at all. We generated a total of 100 bootstrap matrices and the bootstrap values were expressed as percentage of support for each branch.

An IBM cluster of 4 CPUs with 3 GB memory was used for the computation of this study. All the calculations took more than 100 hours.

Analysis of the Subtraction Procedure

In order to elucidate the biological meaning of subtraction we performed an analysis on the example of *E. coli* at string length $K = 5$. There are 1343 887 nonzero 5-strings belonging to 841 832 different string types. Among all the counts the maximal one is 58 for the string "GKSTL". The frequency of the sub-strings "GKST" and "KSTL" is 113 and 77 respectively, while the frequency of the middle string "KST" is 247. Thus the predicted value of "GKSTL" is $\frac{113 \times 77}{247} = 35.2267$ as compare to the real count 58 (neglecting the normalization factor when $L >> K$). The corresponding component in the composition vector after subtraction is $\frac{58 - 35.2267}{35.2267} = 0.646478$.

On the contrary, the string "HAMSC" only appears once in *E. coli*. Its sub-strings "HAMS" and "AMSC" also merely appear once; the frequency of the middle 3-string "AMS" is 198. The predicted value of "HAMSC" is $\frac{1 \times 1}{198} = 0.00505051$. The residual

vector becomes $\frac{1-0.00505051}{0.00505051}=197$, making "HAMSC" the largest component in the vector.

In order to reveal the biological difference between the two strings "GKSTL" and "HAMSC", we searched for the exact match of these two 5-peptides in the Protein Information Resource (PIR) database which contains more than 1.2 million protein sequences at the present time. The string "HAMSC" has 15 matches among which 1 comes from Eukaryotic species, 4 (essentially the same protein) from a virus and 10 from Prokaryotes. Among the 10 Prokaryotes, 4 are from *E. coli* and *Shigella*, 2 from *Samonella*, while all of the 10 are closely related to Enterobacteria. In sharp contrast to "HAMSC", the string "GKSTL" has 6121 matches with proteins of a wide taxonomic assortment from virus to human being. As a commonly occurring 5-peptide, the string "GKSTL" in *E. coli* proteome does not carry much phylogenetic information though it appears most frequently. On the contrary, the 5-peptide "HAMSC" is more characteristic for prokaryotes, especially for Enterobacteria.

Thus frequently occurring strings may not be significant *per se* for inferring phylogenetic relations. In the parlance of classic cladistics they contribute to plesiomorphic characters and should be eliminated in a strict treatment. On the other hand, some strings with small counts may contribute substantially to apomorphic characteristics, if their counts are largely different from what predicted by a reasonable statistical model. The subtraction procedure helps to highlight these significant strings, though it is not always possible to evaluate the effect in a clear-cut way (the above examples are the extreme cases).

The frequency of some peptides in fact is subtracted to zero, though the number of such strings is not large. By counting the number of strings whose values after subtraction fall in the range -0.1 to 0.1, we see they only make a small proportion. It is 6% in *Cyanophor* and 7% for *E. coli*. We cannot say that these zero-strings are not important. Actually they provide necessary information on the degree of dissimilarity among the species which eventually contribute to the classification.

From a mathematical point of view, the subtraction procedure can be considered as removing a multifractal structure before performing a cross-correlation analysis (similar to removing a time-varying mean in time series before computing the cross-correlation of two time series). The multifractal method has been discussed in Anh et al. (2001) and will not be elaborated here.

Results and Discussion

The topologies of the trees generated by distance methods including Fitch-Margoliash (FM), neighbour-joining (NJ) and minimum evolution (ME) are very similar. Fig. 2*a* shows the tree based on ME analysis with bootstrap values from both ME and NJ analyses. Discrepancies of the NJ and FM trees from the ME tree are also shown as alternative topologies in Fig. 2*b*. All the chloroplast genomes form a clade branched in Eubacteria domain and share a most recent common ancestor with cyanobacteria, which is in accordance with the widely accepted endosymbiotic theory that chloroplasts arose from cyanobacteria-like ancestor (Gray 1992, 1999; McFadden 2001b). Apparently, despite massive gene transfer from the endosymbiont to the nucleus of the host cell (Martin and Herrmann 1998; Martin et al. 1998, 2002), our analysis is able to identify

cyanobacteria as the most closely related prokaryotes of chloroplast. We have also attempted to include in our analyses complete genomes of non-photosynthetic plastids of the parasitic flowering plant *Epifagus virginiana* (70 kb), the euglenophyte *Astasia longa* (73 kb) and the apicomplexan *Toxoplasma gondii* (35 kb). All the three taxa appear to be closely related to the two cyanobacteria, with their branches diverged earlier than the other plastids (chloroplasts). We believe such branching positions of the non-photosynthetic plastids are likely to be artifacts (particularly for *Epifagus*, a flowering plant whose plastids have lost all the genes for photosynthesis and chlororespiration, see Wolfe, Morden and Palmer 1992) of massive genome reduction (about 50% or more in the case of apicomplexan) in these degenerate plastids. Thus we have not included these plastids in the tree (Fig. 2). The effect of genome size on the resolving power of our method is under investigation in our laboratory.

Our analysis shows that the chloroplasts are separated into two major clades. One of these corresponds to the green plants *sensu lato*, or chlorophytes *s.l.* (Palmer and Delwiche 1998), which include all taxa with a chlorophyte chloroplast, both primary and secondary endosymbioses in origin. The other clade comprises the glaucophyte *Cyanophora* and members of rhodophytes *s.l.*, which refers to rhodophytes (or red algae) and their secondary symbiotic derivatives, loosely termed chromophytes (including crytophytes, heterokonts, haptophytes and dinoflagellates) (Palmer and Delwiche 1998). The close relationship between *Cyanophora* and rhodophytes *s.l.* agrees with some of the previous analyses (Stirewalt et al. 1995; De Las Rivas, Lozano and Ortiz 2002), although most recent studies suggest that the glaucophyte represents the earliest branch in chloroplast evolution with the green plants *s.l.* and rhodophytes *s.l.* as sister taxa (Martin

et al. 1998, 2002; Stoebe and Kowallik 1999; Adachi et al. 2000; Moreira, Le Guyader and Philippe 2000). Within the rhodophytes *s.l.* clade in our tree (including the two red algae *Cyanidium* and *Porphyra*, the cryptophyte *Guillardia*, and the heterokont *Odontella*), *Porphyra* and *Guillardia* are the most closely related taxa. This agrees with the results from gene clusters comparison between these two species, providing strong evidence that cryptophytes arose by secondary endosymbiosis of a primitive rhodophyte (Douglas and Penny 1999; Stoebe and Kowallik 1999). The paraphyly of *Guillardia* and *Odontella* with respect to the two red algae also suggests independent acquisition of secondary chloroplasts in the heterokont and cryptophyte, in contrast to the hypothesis of a single secondary endosymbiotic event among the chromophytes (Cavelier-Smith 2000). Although a single origin of the chloroplasts in this group is supported in some analyses (De Las Rivas, Lozano and Ortiz 2002; Yoon et al. 2002), the topology of these four taxa in our tree is identical to that based on a recent, traditional analysis of protein-coding genes in the genomes (Martin et al. 2002). Analysis of small subunit ribosomal DNA in the chloroplasts from a wide variety of rhodophytes and chromophytes also indicates that chloroplasts of the latter group have independent origin (Oliveira and Bhattacharya 2000).

The chlorophyte-like chloroplast of euglenophytes is generally believed to have arisen from secondary symbiosis by capture of a green alga in the kinetoplastid lineage (Palmer and Delwiche 1998; Cavelier-Smith 2000). The euglenophyte *Euglena* branches basal to chlorophytes *s.l.* in our tree and is consistent with recent analyses of complete chloroplast genomes (De Las Rivas, Lozano and Ortiz 2002; Martin et al. 2002), although other analyses have placed *Euglena* within the green algae (Van de Peer et al. 1996; Köhler et al. 1997; Turmel, Otis and Lemieux 1999). The chloroplasts of green algae,

including *Chlorella*, *Nephroselmis*, and *Mesostigma*, are more closely related to land plants than to other algae (Wakasugi et al. 1997; Martin et al. 2002). Our analysis however suggests that this assemblage is paraphyletic but the branching order among the three species receives little bootstrap support. ME and NJ trees grouping *Mesostigma* with *Nephroselmis* as prasinophytes are consistent with results from another correlation analysis of complete chloroplast genomes (De Las Rivas, Lozano and Ortiz 2002). Yet an alternate topology (T1) from the MF tree indicates that *Mesostigma* is closely related to the streptophytes (including the charophyte *Chaetosphaeridium* and land plants). Previous molecular phylogenetic studies have also produced conflicting results on the placement of *Mesostigma*. The first complete chloroplast genome analysis of this species showed that it is an ancestral branch of green plant evolution, representing a lineage that emerged before the divergence of green algae and streptophytes (Lemieux, Otis and Turmel 2000). Yet a recent analysis on chloroplast genome sequences showed that it is basal to land plants above the green algae (Martin et al. 2002), in accordance with a multi-gene analysis on a wide variety of charophytes assigning *Mesostigma* to a basal group of charophytes (Karol et al. 2001). The difficulty in resolving the phylogeny of *Mesostigma* in relation to other members of chlorophytes *s.l.* in our analysis is possibly due to the limited taxon sampling of the chloroplasts in green algae and charophytes.

The charophyte *Chaetosphaeridium globosum* represents a basal branch of the streptophyte clade in all analyses. This is consistent with the chloroplast genome analysis of this species (Turmel, Otis and Lemieux 2002), suggesting that charophytes were the immediate ancestor of land plants, or embryophytes (Graham, Cook and Busse 2000). While the support for the angiosperm (flowering plants) clade is strong, its relationships

with other land plants is not well resolved in our analysis. An alternative topology (T2) of both the NJ and FM trees suggests that the angiosperms are more closely related to the liverwort *Marchantia* and the psilophyte *Psilotum* than to the conifer *Pinus*. Interestingly, a recent correlation analysis on the complete chloroplast genomes also indicates the same topology (De Las Rivas, Lozano and Ortiz 2002). Whether this anomaly is due to the almost complete loss of a large inverted repeat in *Pinus* (Wakasugi et al. 1994) as compared to other photosynthetic eukaryotes remains to be investigated. Our analysis clearly separates the angiosperms into two clades corresponding to the monocotyledons and eudicots, the two large clades in current understanding of angiosperm phylogeny (Crane, Friis and Pedersen 1995), although it should be noted that all the monocots included in the tree are members of a single family (Poaceae). The branching order within each clade is not well supported by bootstrapping. A different topology (T3) among three of the eudicots (*Spinacia*, *Nicotiana* and *Arabidopsis*) is suggested by the both the NJ and FM trees as compared to the ME tree.

Our simple correlation analysis on the complete chloroplast genomes has yielded a tree that is in good agreement with our current knowledge on the origin of the chloroplasts and the phylogenetic relationships of different groups of photosynthetic eukaryotes as elucidated previously by traditional analyses of the chloroplast genomes and other molecular/ultrastructural approaches (e.g., Martin et al. 2002; De Las Rivas, Lozano and Ortiz 2002; see also Palmer and Delwiche 1998, McFdden 2001a,b for reviews).

Remark 2. Removal of the random background has been an essential step in our approach. The phylogenetic results are quite different without this procedure. In fact, without this removal, the topology becomes worse. Qi, Wang and Hao (private

communication) generated a tree of 109 species without this removal. In the Kingdom level, Archaea, Bacteria and Eukaryotes are mixed together, and are not clearly divided into three groups as in the tree of Figure 1 of Qi et al. (in press). The classification in the middle level in the tree without removal is not satisfactory as it cannot provide enough meaningful information to compare with biological classification. Only in the lower level that some small groups are consistent with existing biological results. We also generated the chloroplast tree developed without removal of random background. The tree shows that, although the grouping of chloroplast is obtained, the Archaea and Bacteria are still mixed.

Our approach circumvents the ambiguity in the selection of genes from complete genomes for phylogenetic reconstruction, and is also faster than the traditional approaches of phylogenetic analysis, particularly when dealing with a large number of genomes. Moreover, since multiple sequence alignment is not necessary, the intrinsic problems associated with this complex procedure can be avoided. In contrast to a recent similar analysis on mitochondrial genomes based on compositional vector (Stuart, Moffet and Baker 2002; Stuart, Moffet and Leader 2002), our approach does not require prior information on gene families in the genome and is also simpler in the method used for subtraction of random background from the data set (see Materials and Methods). We have also shown that this approach is applicable for analyzing the much larger genomes of chloroplast, as well as the prokaryotes (Qi, Wang and Hao, in press). We believe that the present approach is an important step towards the analysis of the wealth of information provided by genome projects. In view of the lower resolving power (i.e., relatively low bootstrap support in most of the branches) as compared to the conventional analysis of chloroplast genomes (e.g., Martin et al. 2002), further refinements of the method is being explored in our laboratories, along with the question on the nature of the

phylogenetic signals revealed in our method. It is hoped that efforts in this line of research will provide us with fast and useful tools in comparative genome analysis as well as insights on genome structure and evolution.

Acknowledgments

We thank C.P. Li and K.C. Cheung for technical assistance, B.-L. Hao for discussion, and C.K. Wong for comments on the draft manuscript. Financial support was provided by an AoE Fund of The Chinese University of Hong Kong (K.H. Chu), Youth Foundation of the Chinese National Natural Science Foundation (grant no. 10101022), and Postdoctoral Research Support Grant (no. 9900658) of Queensland University of Technology (Z.-G. Yu). The use of the 64 CPU IBM Cluster at Peking University and the facilities of Centre of Theoretical Biology of Fudan University are gratefully acknowledged.

LITERATURE CITED

ADACHI, J., P. J. WADDELL, W. MARTIN, and M. HASEGAWA. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. **50**:348-358.

ANH, V.V., K.S. LAU, and Z.G. YU 2001. Multifractal characterization of complete genomes. J. Phys. A: Math. Gen. **34**: 7127-7139.

BRENDEL, V., J. S. BECKMANN, and E. N. TRIFONOV. 1986. Linguistics of nucleotide sequences: Morphology and comparison of vocabularies, J. Biomol. Struct. Dyn. **4**:11-21.

CAVALIER-SMITH, T. 2000. Membrane heredity and early chloroplast evolution. Trends Plant Sci. **5**:174-182.

CRANE, P. R., E. M. FRIIS, and K. R. PEDERSEN. 1995. The origin and early diversification of angiosperms. Nature **374**:27-33.

DE LAS RIVAS, J., J. J. LOZANO, and A. R. ORTIZ. 2002. Comparative analysis of chloroplast genomes: Functional annotation, genome-based phylogeny, and deduced evolutionary patterns. Genome Res. **12**:567-583.

DELWICHE, C. F. 1999. Tracing the thread of plastid diversity through the tapestry of life. Am. Nat. **154**:S164-S177.

DOUGLAS, S. E., and S. L. PENNY. 1999. The plastid genome of the cryptophyte alga, *Guillardia theta*: Complete sequence and conserved synteny groups confirm its common ancestry with red algae. J. Mol. Evol. **48**:236-244.

EDWARDS, S. V., B. FERTIL, A. GIRON, and P. J. DESCHAVANNE. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. Syst. Biol. **51**:599-613.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science 1**55**:279-284.

FITZ-GIBBON, S. T., and C. H. HOUSE. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. **27**:4218-4222.

GRAHAM, L. E., M. E. COOK, and J. E. BUSSE. 2000. The origin of plants: Body plan changes contributing to a major evolutionary radiation. Proc. Natl. Acad. Sci. U.S.A. **97**:4535-4540.

GRAY, M. W. 1992. The endosymbiont hypothesis revisited. Int. Rev. Cytol. **141**:233-357.

GRAY, M. W. 1999. Evolution of organellar genomes. Curr. Opin. Genet. Dev. **9**:678-687.

KAROL, K. G., R. M. MCCOURT, M. T. CIMINO, and C. F. DELWICHE. 2001. The closest living relatives of land plants. Science **294**:2351-2353.

KÖHLER, S., C. F. DELWICHE, P. W. DENNY, L. G. TILNEY, P. WEBSTER, R. J. M. WILSON, J. D. PALMER, and D. S. ROOS. 1997. A plastid of probable green algal origin in apicomplexan parasites. Science **275**:1485-1489.

LEMIEUX, C., C. OTIS, and M. TURMEL. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature **403**:649-652.

LI, M., J. H. BADGER, X. CHEN, S. KWONG, P. KEARNEY, and H. ZHANG. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics **17**:149-154.

LIN, J., and M. GERSTEIN. 2000. Whole-genome trees based on the occurrence of folds and orthologs, implications for comparing genomes at different levels. Genome Res. **10**:808-818.

MCFADDEN, G. I. 2001a. Primary and secondary endosymbiosis and the origin of plastids. J. Phycol. **37**:951-959.

MCFADDEN, G. I. 2001b. Chloroplast origin and integration. Plant Physiol. **125**:50-53.

MARTIN, W., and R. G. HERRMANN. 1998. Gene transfer from organelles to the nucleus: How much, what happens, and why? Plant Physiol. **118**:9-17.

MARTIN, W., B. STOEBE, V. GOREMYKIN, S. HANSMANN, M. HASEGAWA, and K. V. KOWALLIK. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature **393**:162-165.

MARTIN, W., T. RUJAN, E. RICHLY, A. HANSEN, S. CORNELSEN, T. LINS, D. LEISTER, B. STOEBE, M. HASEGAWA, and D. PENNY. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. U.S.A. **99**:12246-12251.

MOREIRA, D., H. LE GUYADER, and H. PHILIPPE. 2000. The origin of red algae and the evolution of chloroplasts. Nature **405**:69-72.

OLIVEIRA, M. C., and D. BHATTACHARYA. 2000. Phylogeny of the Bangiophycidae (Rhodophyta) and the secondary endosymbiotic origin of algal plastids. Am. J. Bot. **87**:482-492.

PALMER, J. D., and C. F. DELWICHE. 1998. The origin and evolution of plastids and their genomes. In Molecular Systematics of Plants II DNA Sequencing (eds. Soltis, D.E., Soltis, P.S. and Doyle, J.J.), pp. 345-409. Kluwer, London.

PERCUS, J.K. 2002. Mathematics of Genome Analysis. Cambridge University Press.

QI, J., B. WANG, and B. HAO. (in press). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406-425.

SAITOU, N., and T. IMANISHI. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. **6**:514-525.

SANKOFF, D., G. LEADUC, N. ANTOINE, B. PAQUIN, B. F. LANG, and R. CEDERGREN. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. U.S.A. **89**:6575-6579.

STIREWALT, V. L., C. B. MICHALOWSKI, W. LOFFELHARDT, H. J. BOHNERT, and D. A. BRYANT. 1995. Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. Plant Mol. Biol. Rep. **13**:327-332.

STOEBE, B., and K. V. KOWALLIK. 1999. Gene-cluster analysis in chloroplast genomics. Genome Analysis Outlook **15**:344-347.

STUART, G. W., K. MOFFET, and S. BAKER. 2002. Integrated gene species phylogenies from unaligned whole genome protein sequences. Bioinformatics **18**:100-108.

STUART, G. W., K. MOFFET, and J. J. LEADER. 2002. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Mol. Biol. Evol. **19**:554-562.

TEKAIA, F., A. LAZCANO, and B. DUJON. 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. **9**:550-557.

TURMEL, M., C. OTIS, and C. LEMIEUX. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: Insights into the architecture of ancestral chloroplast genomes. Proc. Natl. Acad. Sci. U.S.A. **96**:10248-10253.

TURMEL, M., C. OTIS, and C. LEMIEUX. 2002. The chloroplast and mitochondrial genome sequences of the charophyte C*haetosphaeridium globosum:* Insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc. Natl. Acad. Sci. U.S.A. **99**:11275-11280.

VAN dE PEER, Y., S. A. RENSING, U. G. MAIER, and R. DE WACHTER. 1996. Substitution

rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc. Natl. Acad. Sci. U.S.A. **93**:7732-7736.

WAKASUGI, T., J. TSUDZUKI, S. ITO, K. NAKASHIMA, T. TSUDZUKI, and M. SUGIURA. 1994. Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc. Natl. Acad. Sci. U.S.A. **91**:9794-9798.

WAKASUGI, T., T. NAGAI, M. KAPOOR, M. SUGITA, M. ITO, S. ITO, J. TSUDZUKI, K. NAKASHIMA, T. TSUDZUKI, Y. SUZUKI, A. HAMADA, T. OHTA, A. INAMURA, K. YOSHINAGA, and M. SUGIURA. 1997. Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: The existence of genes possibly involved in chloroplast division. Proc. Natl. Acad. Sci. U.S.A. **94**:5967-5972.

WEISS, O., M. A. JIMENEZ, and H. HERZEL. 2000. Information content of protein sequences. J. Theor. Biol. **206**:379-386.

WOLFE, K. H., C. W. MORDEN, and J. D. PALMER. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. U.S.A. **89**:10648-10652.

YOON, H. S., J. D. HACKETT, G. PINTO, and D. BHATTACHARYA. 2002. The single, ancient origin of chromist plastids. Proc. Natl. Acad. Sci. U.S.A. **99**:15507-15512.

YU, Z.-G., and P. JIANG. 2001. Distance, correlation and mutual information among portraits of organisms based on complete genomes. Phys. Lett. A **286**:34-46.

Figure legend:

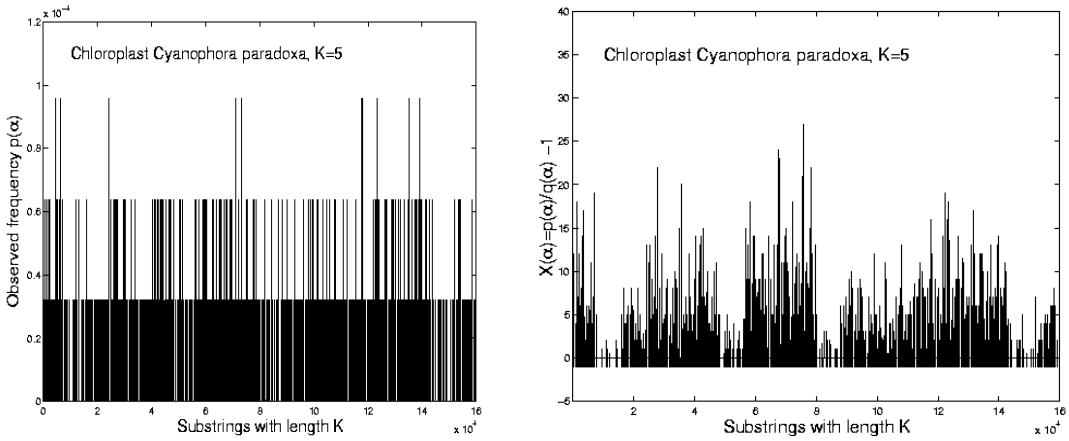

Fig. 1. A segment of *p* for Chloroplast *Cyanophora paradoxa* in the left figure and the corresponding sequence *X* for the same set of *K*-strings in the right figure.

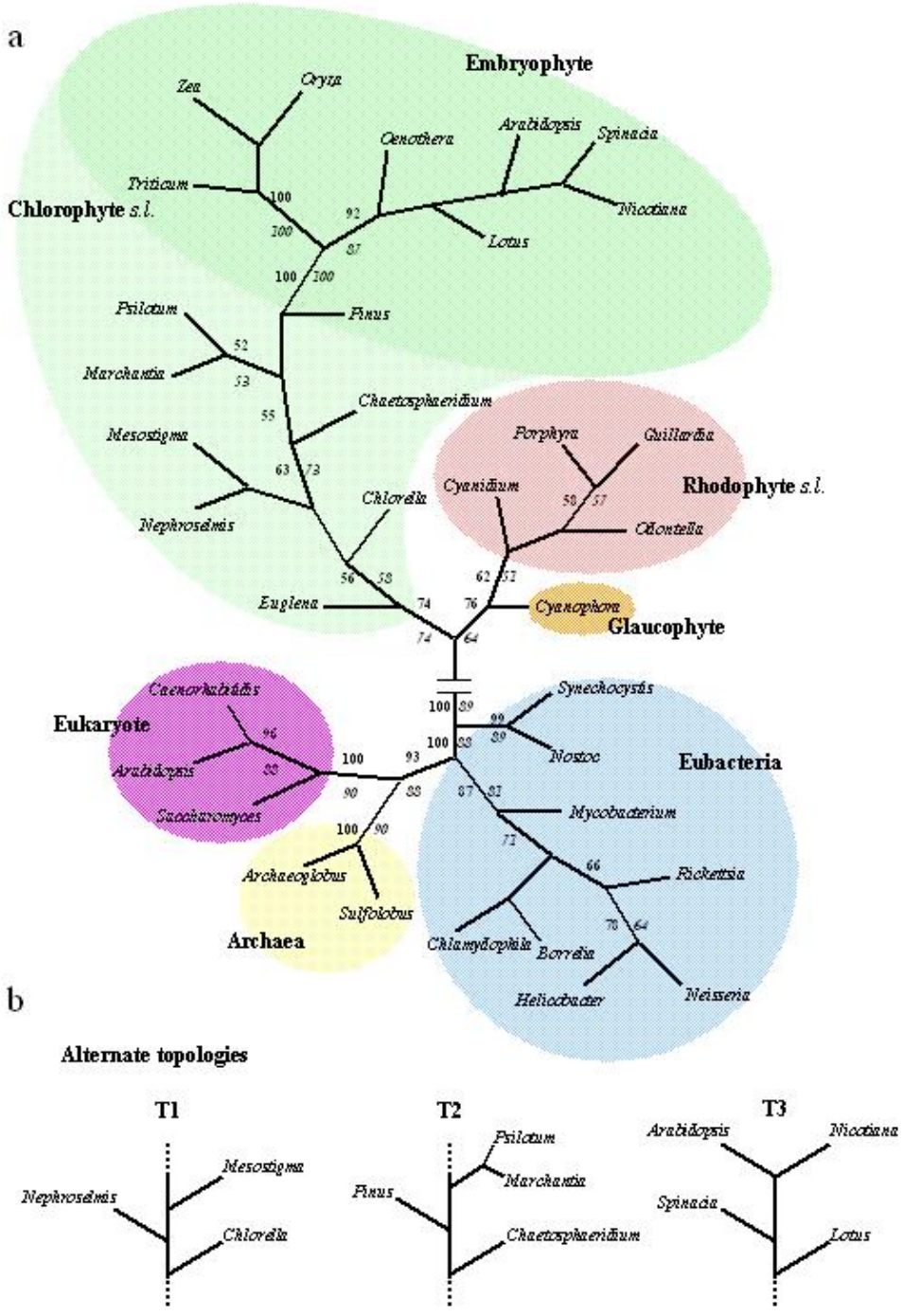
a
Embryophyte
Zea
Oryza
Oenothera
Arabidopsis
Spinacia
Triticum
Nicotiana
Chlorophyte s.l.
Lotus
Psilotum
Pinus
Marchantia
Chaetosphaeridium
Mesostigma
Porphyra
Guillardia
Cyanidium
Rhodophyte s.l.
Chlorella
Nephroselmis
Cyanidium
Ollontella
Euglena
Cyanophora
Glaucophyte
Caenorhabditis
Synechocystis
Eukaryote
Arabidopsis
Nostoc
Eubacteria
Saccharomyces
Mycobacterium
Archaeoglobus
Rickettsia
Sulfolobus
Archaea
Chlamydophila
Borrelia
Heliobacter
Neisseria
b
Alternate topologies
T1
Mesostigma
Nephroselmis
Chlorella
T2
Psilotum
Marchantia
Pinus
Chaetosphaeridium
T3
Arabidopsis
Nicotiana
Spinacia
Lotus

Fig. 2 Phylogeny of chloroplast genomes based on correlation analysis. **a,** Topology of chloroplast genomes together with selected genomes from eubacteria, archaea, and eukaryotes using minimum evolution (ME) analysis. The numbers on each branch show the bootstrap support (100 replicates) based on ME and neighbour-joining (NJ, in italic) analyses. Values <50 are not shown. Values shown among the eubacteria, archaea, and eukaryotes are based on the analysis of all 34 genomes. Values shown among the chloroplasts are based on analysis of these 21 genomes using *Synechorcystis* as outgroup. **b,** Alternative topologies of the trees based on Fitch-Margoliash (FM, for T1), or both FM and NJ (for T2 and T3) analyses.